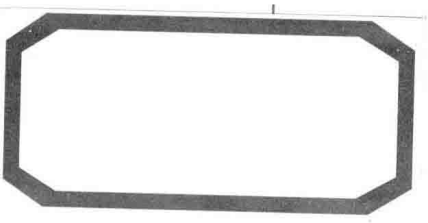
Python机器学习（影印版）

Python Machine Learning

Sebastian Raschka 著

[PACKT] PUBLISHING

东南大学出版社
SOUTHEAST UNIVERSITY PRESS



Python 机器学习(影印版)

Sebastian Raschka 著

南京　东南大学出版社

图书在版编目(CIP)数据

Python 机器学习:英文/(美)塞巴斯蒂安·拉什卡(Sebastian Raschka)著. —影印本. —南京:东南大学出版社,2017.4

书名原文:Python Machine Learning

ISBN 978-7-5641-7077-6

Ⅰ.①P… Ⅱ.①塞… Ⅲ.①软件工具—程序设计—英文 Ⅳ.①TP311.561

中国版本图书馆 CIP 数据核字(2017)第 051740 号

Python 机器学习(影印版)

出版发行:东南大学出版社
地　　址:南京四牌楼 2 号　　邮编:210096
出 版 人:江建中
网　　址:http://www.seupress.com
电子邮件:press@seupress.com
印　　刷:常州市武进第三印刷有限公司
开　　本:787 毫米×980 毫米　　16 开本
印　　张:28.25
字　　数:553 千字
版　　次:2017 年 4 月第 1 版
印　　次:2017 年 4 月第 1 次印刷
书　　号:ISBN 978-7-5641-7077-6
定　　价:87.00 元

本社图书若有印装质量问题,请直接与营销部联系。电话(传真):025-83791830

Credits

Author

Sebastian Raschka

Reviewers

Richard Dutton

Dave Julian

Vahid Mirjalili

Hamidreza Sattari

Dmytro Taranovsky

Commissioning Editor

Akram Hussain

Acquisition Editors

Rebecca Youe

Meeta Rajani

Content Development Editor

Riddhi Tuljapurkar

Technical Editors

Madhunikita Sunil Chindarkar

Taabish Khan

Copy Editors

Roshni Banerjee

Stephan Copestake

Project Coordinator

Kinjal Bari

Proofreader

Safis Editing

Indexer

Hemangini Bari

Graphics

Sheetal Aute

Abhinash Sahu

Production Coordinator

Shantanu N. Zagade

Cover Work

Shantanu N. Zagade

Foreword

We live in the midst of a data deluge. According to recent estimates, 2.5 quintillion (10^{18}) bytes of data are generated on a daily basis. This is so much data that over 90 percent of the information that we store nowadays was generated in the past decade alone. Unfortunately, most of this information cannot be used by humans. Either the data is beyond the means of standard analytical methods, or it is simply too vast for our limited minds to even comprehend.

Through Machine Learning, we enable computers to process, learn from, and draw actionable insights out of the otherwise impenetrable walls of big data. From the massive supercomputers that support Google's search engines to the smartphones that we carry in our pockets, we rely on Machine Learning to power most of the world around us—often, without even knowing it.

As modern pioneers in the brave new world of big data, it then behooves us to learn more about Machine Learning. What is Machine Learning and how does it work? How can I use Machine Learning to take a glimpse into the unknown, power my business, or just find out what the Internet at large thinks about my favorite movie? All of this and more will be covered in the following chapters authored by my good friend and colleague, Sebastian Raschka.

When away from taming my otherwise irascible pet dog, Sebastian has tirelessly devoted his free time to the open source Machine Learning community. Over the past several years, Sebastian has developed dozens of popular tutorials that cover topics in Machine Learning and data visualization in Python. He has also developed and contributed to several open source Python packages, several of which are now part of the core Python Machine Learning workflow.

Owing to his vast expertise in this field, I am confident that Sebastian's insights into the world of Machine Learning in Python will be invaluable to users of all experience levels. I wholeheartedly recommend this book to anyone looking to gain a broader and more practical understanding of Machine Learning.

Dr. Randal S. Olson
Artificial Intelligence and Machine Learning Researcher, University of Pennsylvania

About the Author

Sebastian Raschka is a PhD student at Michigan State University, who develops new computational methods in the field of computational biology. He has been ranked as the number one most influential data scientist on GitHub by Analytics Vidhya. He has a yearlong experience in Python programming and he has conducted several seminars on the practical applications of data science and machine learning. Talking and writing about data science, machine learning, and Python really motivated Sebastian to write this book in order to help people develop data-driven solutions without necessarily needing to have a machine learning background.

He has also actively contributed to open source projects and methods that he implemented, which are now successfully used in machine learning competitions, such as Kaggle. In his free time, he works on models for sports predictions, and if he is not in front of the computer, he enjoys playing sports.

I would like to thank my professors, Arun Ross and Pang-Ning Tan, and many others who inspired me and kindled my great interest in pattern classification, machine learning, and data mining.

I would like to take this opportunity to thank the great Python community and developers of open source packages who helped me create the perfect environment for scientific research and data science.

A special thanks goes to the core developers of scikit-learn. As a contributor to this project, I had the pleasure to work with great people, who are not only very knowledgeable when it comes to machine learning, but are also excellent programmers.

Lastly, I want to thank you all for showing an interest in this book, and I sincerely hope that I can pass on my enthusiasm to join the great Python and machine learning communities.

About the Reviewers

Richard Dutton started programming the ZX Spectrum when he was 8 years old and his obsession carried him through a confusing array of technologies and roles in the fields of technology and finance.

He has worked with Microsoft, and as a Director at Barclays, his current obsession is a mashup of Python, machine learning, and block chain.

If he's not in front of a computer, he can be found in the gym or at home with a glass of wine while he looks at his iPhone. He calls this balance.

Dave Julian is an IT consultant and teacher with over 15 years of experience. He has worked as a technician, project manager, programmer, and web developer. His current projects include developing a crop analysis tool as part of integrated pest management strategies in greenhouses. He has a strong interest in the intersection of biology and technology with a belief that smart machines can help solve the world's most important problems.

Vahid Mirjalili received his PhD in mechanical engineering from Michigan State University, where he developed novel techniques for protein structure refinement using molecular dynamics simulations. Combining his knowledge from the fields of statistics, data mining, and physics he developed powerful data-driven approaches that helped him and his research group to win two recent worldwide competitions for protein structure prediction and refinement, CASP, in 2012 and 2014.

While working on his doctorate degree, he decided to join the Computer Science and Engineering Department at Michigan State University to specialize in the field of machine learning. His current research projects involve the development of unsupervised machine learning algorithms for the mining of massive datasets. He is also a passionate Python programmer and shares his implementations of clustering algorithms on his personal website at `http://vahidmirjalili.com`.

Hamidreza Sattari is an IT professional and has been involved in several areas of software engineering, from programming to architecture, as well as management. He holds a master's degree in software engineering from Herriot-Watt University, UK, and a bachelor's degree in electrical engineering (electronics) from Tehran Azad University, Iran. In recent years, his areas of interest have been big data and Machine Learning. He coauthored the book *Spring Web Services 2 Cookbook* and he maintains his blog at `http://justdeveloped-blog.blogspot.com/`.

Dmytro Taranovsky is a software engineer with an interest and background in Python, Linux, and machine learning. Originally from Kiev, Ukraine, he moved to the United States in 1996. From an early age, he displayed a passion for science and knowledge, winning mathematics and physics competitions. In 1999, he was chosen to be a member of the U.S. Physics Team. In 2005, he graduated from the Massachusetts Institute of Technology, majoring in mathematics. Later, he worked as a software engineer on a text transformation system for computer-assisted medical transcriptions (eScription). Although he originally worked on Perl, he appreciated the power and clarity of Python, and he was able to scale the system to very large data sizes. Afterwards, he worked as a software engineer and analyst for an algorithmic trading firm. He also made significant contributions to the foundation of mathematics, including creating and developing an extension to the language of set theory and its connection to large cardinal axioms, developing a notion of constructive truth, and creating a system of ordinal notations and implementing them in Python. He also enjoys reading, likes to go outdoors, and tries to make the world a better place.

www.PacktPub.com

Support files, eBooks, discount offers, and more

For support files and downloads related to your book, please visit `www.PacktPub.com`.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at `www.PacktPub.com` and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at `service@packtpub.com` for more details.

At `www.PacktPub.com`, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.

`https://www2.packtpub.com/books/subscription/packtlib`

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can search, access, and read Packt's entire library of books.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Free access for Packt account holders

If you have an account with Packt at `www.PacktPub.com`, you can use this to access PacktLib today and view 9 entirely free books. Simply use your login credentials for immediate access.

Preface

I probably don't need to tell you that machine learning has become one of the most exciting technologies of our time and age. Big companies, such as Google, Facebook, Apple, Amazon, IBM, and many more, heavily invest in machine learning research and applications for good reasons. Although it may seem that machine learning has become the buzzword of our time and age, it is certainly not a hype. This exciting field opens the way to new possibilities and has become indispensable to our daily lives. Talking to the voice assistant on our smart phones, recommending the right product for our customers, stopping credit card fraud, filtering out spam from our e-mail inboxes, detecting and diagnosing medical diseases, the list goes on and on.

If you want to become a machine learning practitioner, a better problem solver, or maybe even consider a career in machine learning research, then this book is for you! However, for a novice, the theoretical concepts behind machine learning can be quite overwhelming. Yet, many practical books that have been published in recent years will help you get started in machine learning by implementing powerful learning algorithms. In my opinion, the use of practical code examples serve an important purpose. They illustrate the concepts by putting the learned material directly into action. However, remember that with great power comes great responsibility! The concepts behind machine learning are too beautiful and important to be hidden in a black box. Thus, my personal mission is to provide you with a different book; a book that discusses the necessary details regarding machine learning concepts, offers intuitive yet informative explanations on how machine learning algorithms work, how to use them, and most importantly, how to avoid the most common pitfalls.

If you type "machine learning" as a search term in Google Scholar, it returns an overwhelmingly large number-1,800,000 publications. Of course, we cannot discuss all the nitty-gritty details about all the different algorithms and applications that have emerged in the last 60 years. However, in this book, we will embark on an exciting journey that covers all the essential topics and concepts to give you a head start in this field. If you find that your thirst for knowledge is not satisfied, there are many useful resources that can be used to follow up on the essential breakthroughs in this field.

If you have already studied machine learning theory in detail, this book will show you how to put your knowledge into practice. If you have used machine learning techniques before and want to gain more insight into how machine learning really works, this book is for you! Don't worry if you are completely new to the machine learning field; you have even more reason to be excited. I promise you that machine learning will change the way you think about the problems you want to solve and will show you how to tackle them by unlocking the power of data.

Before we dive deeper into the machine learning field, let me answer your most important question, "why Python?" The answer is simple: it is powerful yet very accessible. Python has become the most popular programming language for data science because it allows us to forget about the tedious parts of programming and offers us an environment where we can quickly jot down our ideas and put concepts directly into action.

Reflecting on my personal journey, I can truly say that the study of machine learning made me a better scientist, thinker, and problem solver. In this book, I want to share this knowledge with you. Knowledge is gained by learning, the key is our enthusiasm, and the true mastery of skills can only be achieved by practice. The road ahead may be bumpy on occasions, and some topics may be more challenging than others, but I hope that you will embrace this opportunity and focus on the reward. Remember that we are on this journey together, and throughout this book, we will add many powerful techniques to your arsenal that will help us solve even the toughest problems the data-driven way.

What this book covers

Chapter 1, Giving Computers the Ability to Learn from Data, introduces you to the main subareas of machine learning to tackle various problem tasks. In addition, it discusses the essential steps for creating a typical machine learning model building pipeline that will guide us through the following chapters.

Chapter 2, Training Machine Learning Algorithms for Classification, goes back to the origin of machine learning and introduces binary perceptron classifiers and adaptive linear neurons. This chapter is a gentle introduction to the fundamentals of pattern classification and focuses on the interplay of optimization algorithms and machine learning.

Chapter 3, A Tour of Machine Learning Classifiers Using Scikit-learn, describes the essential machine learning algorithms for classification and provides practical examples using one of the most popular and comprehensive open source machine learning libraries, scikit-learn.

Chapter 4, Building Good Training Sets – Data Preprocessing, discusses how to deal with the most common problems in unprocessed datasets, such as missing data. It also discusses several approaches to identify the most informative features in datasets and teaches you how to prepare variables of different types as proper inputs for machine learning algorithms.

Chapter 5, Compressing Data via Dimensionality Reduction, describes the essential techniques to reduce the number of features in a dataset to smaller sets while retaining most of their useful and discriminatory information. It discusses the standard approach to dimensionality reduction via principal component analysis and compares it to supervised and nonlinear transformation techniques.

Chapter 6, Learning Best Practices for Model Evaluation and Hyperparameter Tuning, discusses the do's and don'ts for estimating the performances of predictive models. Moreover, it discusses different metrics for measuring the performance of our models and techniques to fine-tune machine learning algorithms.

Chapter 7, Combining Different Models for Ensemble Learning, introduces you to the different concepts of combining multiple learning algorithms effectively. It teaches you how to build ensembles of experts to overcome the weaknesses of individual learners, resulting in more accurate and reliable predictions.

Chapter 8, Applying Machine Learning to Sentiment Analysis, discusses the essential steps to transform textual data into meaningful representations for machine learning algorithms to predict the opinions of people based on their writing.

Chapter 9, Embedding a Machine Learning Model into a Web Application, continues with the predictive model from the previous chapter and walks you through the essential steps of developing web applications with embedded machine learning models.

Chapter 10, Predicting Continuous Target Variables with Regression Analysis, discusses the essential techniques for modeling linear relationships between target and response variables to make predictions on a continuous scale. After introducing different linear models, it also talks about polynomial regression and tree-based approaches.

Chapter 11, Working with Unlabeled Data – Clustering Analysis, shifts the focus to a different subarea of machine learning, unsupervised learning. We apply algorithms from three fundamental families of clustering algorithms to find groups of objects that share a certain degree of similarity.

Chapter 12, Training Artificial Neural Networks for Image Recognition, extends the concept of gradient-based optimization, which we first introduced in *Chapter 2, Training Machine Learning Algorithms for Classification*, to build powerful, multilayer neural networks based on the popular backpropagation algorithm.

Chapter 13, Parallelizing Neural Network Training with Theano, builds upon the knowledge from the previous chapter to provide you with a practical guide for training neural networks more efficiently. The focus of this chapter is on Theano, an open source Python library that allows us to utilize multiple cores of modern GPUs.

What you need for this book

The execution of the code examples provided in this book requires an installation of Python 3.4.3 or newer on Mac OS X, Linux, or Microsoft Windows. We will make frequent use of Python's essential libraries for scientific computing throughout this book, including SciPy, NumPy, scikit-learn, matplotlib, and pandas.

The first chapter will provide you with instructions and useful tips to set up your Python environment and these core libraries. We will add additional libraries to our repertoire and installation instructions are provided in the respective chapters: the NLTK library for natural language processing (*Chapter 8, Applying Machine Learning to Sentiment Analysis*), the Flask web framework (*Chapter 9, Embedding a Machine Learning Algorithm into a Web Application*), the seaborn library for statistical data visualization (*Chapter 10, Predicting Continuous Target Variables with Regression Analysis*), and Theano for efficient neural network training on graphical processing units (*Chapter 13, Parallelizing Neural Network Training with Theano*).

Who this book is for

If you want to find out how to use Python to start answering critical questions of your data, pick up *Python Machine Learning*—whether you want to start from scratch or want to extend your data science knowledge, this is an essential and unmissable resource.

Conventions

In this book, you will find a number of text styles that distinguish between different kinds of information. Here are some examples of these styles and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows: "And already installed packages can be updated via the `--upgrade` flag."

A block of code is set as follows:

```
>>> import matplotlib.pyplot as plt
>>> import numpy as np

>>> y = df.iloc[0:100, 4].values
>>> y = np.where(y == 'Iris-setosa', -1, 1)
>>> X = df.iloc[0:100, [0, 2]].values
>>> plt.scatter(X[:50, 0], X[:50, 1],
...             color='red', marker='x', label='setosa')
>>> plt.scatter(X[50:100, 0], X[50:100, 1],
...             color='blue', marker='o', label='versicolor')
>>> plt.xlabel('petal length')
>>> plt.ylabel('sepal length')
>>> plt.legend(loc='upper left')
>>> plt.show()
```

Any command-line input or output is written as follows:

```
> dot -Tpng tree.dot -o tree.png
```

New terms and **important words** are shown in bold. Words that you see on the screen, for example, in menus or dialog boxes, appear in the text like this: "After we click on the **Dashboard** button in the top-right corner, we have access to the control panel shown at the top of the page."

Warnings or important notes appear in a box like this.

Tips and tricks appear like this.

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book—what you liked or disliked. Reader feedback is important for us as it helps us develop titles that you will really get the most out of.

To send us general feedback, simply e-mail `feedback@packtpub.com`, and mention the book's title in the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide at `www.packtpub.com/authors`.

Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you to get the most from your purchase.

Downloading the example code

You can download the example code files from your account at `http://www.packtpub.com` for all the Packt Publishing books you have purchased. If you purchased this book elsewhere, you can visit `http://www.packtpub.com/support` and register to have the files e-mailed directly to you.

Errata

Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you find a mistake in one of our books—maybe a mistake in the text or the code—we would be grateful if you could report this to us. By doing so, you can save other readers from frustration and help us improve subsequent versions of this book. If you find any errata, please report them by visiting `http://www.packtpub.com/submit-errata`, selecting your book, clicking on the **Errata Submission Form** link, and entering the details of your errata. Once your errata are verified, your submission will be accepted and the errata will be uploaded to our website or added to any list of existing errata under the Errata section of that title.

To view the previously submitted errata, go to `https://www.packtpub.com/books/content/support` and enter the name of the book in the search field. The required information will appear under the **Errata** section.

Piracy

Piracy of copyrighted material on the Internet is an ongoing problem across all media. At Packt, we take the protection of our copyright and licenses very seriously. If you come across any illegal copies of our works in any form on the Internet, please provide us with the location address or website name immediately so that we can pursue a remedy.

Please contact us at `copyright@packtpub.com` with a link to the suspected pirated material.

We appreciate your help in protecting our authors and our ability to bring you valuable content.

Questions

If you have a problem with any aspect of this book, you can contact us at `questions@packtpub.com`, and we will do our best to address the problem.

[PACKT] open source
community experience distilled
PUBLISHING

Thank you for buying
Python Machine Learning

About Packt Publishing

Packt, pronounced 'packed', published its first book, *Mastering phpMyAdmin for Effective MySQL Management*, in April 2004, and subsequently continued to specialize in publishing highly focused books on specific technologies and solutions.

Our books and publications share the experiences of your fellow IT professionals in adapting and customizing today's systems, applications, and frameworks. Our solution-based books give you the knowledge and power to customize the software and technologies you're using to get the job done. Packt books are more specific and less general than the IT books you have seen in the past. Our unique business model allows us to bring you more focused information, giving you more of what you need to know, and less of what you don't.

Packt is a modern yet unique publishing company that focuses on producing quality, cutting-edge books for communities of developers, administrators, and newbies alike. For more information, please visit our website at `www.packtpub.com`.

About Packt Open Source

In 2010, Packt launched two new brands, Packt Open Source and Packt Enterprise, in order to continue its focus on specialization. This book is part of the Packt Open Source brand, home to books published on software built around open source licenses, and offering information to anybody from advanced developers to budding web designers. The Open Source brand also runs Packt's Open Source Royalty Scheme, by which Packt gives a royalty to each open source project about whose software a book is sold.

Writing for Packt

We welcome all inquiries from people who are interested in authoring. Book proposals should be sent to `author@packtpub.com`. If your book idea is still at an early stage and you would like to discuss it first before writing a formal book proposal, then please contact us; one of our commissioning editors will get in touch with you.

We're not just looking for published authors; if you have strong technical skills but no writing experience, our experienced editors can help you develop a writing career, or simply get some additional reward for your expertise.